Department of Computer Science and Engineering, Agneta Nilsson Software Engineering Master Thesis Proposal, 30hec

<Title of your thesis work>

|  |  |
| --- | --- |
| *Helena Ólafsdóttir:* | *Completed courses relevant for thesis work:*  *DAT231 Empirical Software Engineering*  *TDA231 Algorithms for machine learning and inference*  *DAT345 Techniques for Large-scale Data*  *DAT340 Applied Machine Learning*  *TIN175 Introduction to Artificial intelligence* |

# Introduction

An introduction provides readers with the background information for the research pro- posed (or reported in the paper), with the purpose to provide an understanding of how the research is related to other research (Wilkinson 1991). In an introduction, the writer should (Creswell 2002):

* + Create reader interest in the topic
  + Lay the broad foundation for the problem that leads to the study
  + Place the study within the larger context of the scholarly literature
  + Reach out to a specific audience

An enormous amount of information hides within the walls of big data. New ways of pulling out this information are constantly being discovered and used to make more intelligent and better informed descisions. Machine learning plays a key role in this development and more and more fields are introducing machine learning techniques to optimize their procedures. Retail is no exception with machine learning models playing a big role in the improvements of sales forecasts, market basket analysis, and more.

Promotional planning in practice is often only based on simplistic information like stock status or even just the retailer's intuition. A promotion of a product however has complex effects on the sale of not just the promoted product but other products as well. Information about these effects are available in the sale history and this data can be used to make better decisions based on previous experience. By taking a look at past promotions it is possible to see how promotions of particular products affect the sales of other products and therefore identify both correlated and cannibalization products. Providing the retailer with information about the highly correlated/cannibalized products will help him make an informed decision give him a better estimate of the overall profit of the promotion.

*\* add something about how this is related to other research??*

# Statement of the problem

The statement of the problem is the foundation for the construction of any research pro- posal. In addition to being an integral part of selecting a research topic, it also helps to se- lect research design. It serves as the bases for determining research objectives, formulation of research hypotheses or research questions, and planning the research design (Booth et al 2003). It allows the researcher to describe the problem systematically, to reflect on its im- portance, its priority and to point out why the proposed research on the problem should be undertaken.

A problem might be defined as the issue that exists in the literature, theory, or practice that leads to a need for the study. It is important in a proposal that the problem stands out and that the reader can easily recognize it.

* + A problem statement should be presented within a context, and that context should be provided and briefly explained, including a discussion of the conceptual or the- oretical framework in which it is embedded.
  + Clearly and succinctly identify and explain the problem within the framework of the theory or line of inquiry that supports the study.
  + State the problem in terms intelligible to someone who is generally sophisticated but who is relatively uninformed in the area of your investigation.

Effective problem statements answer the question: Why does this research need to be con- ducted? If the writer is unable to answer this question clearly and succinctly, the statement

of the problem will be perceived as vague and diffuse.

It is a known effect of recommender models that the recommendation of one product can have both positive and negative effect on non-recommended products. Promotions in retail have similar effects. The promotion of a product can cause an increase in sales of positively correlated products whereas negatively correlated (cannibalization) products can experience decrease in sales. Without knowing what the correlated and cannibalization products are, it is not possible for a retailer to make a fully informed decision when planning a promotion. By giving him this information, he should however be able to know the overall expected profit of promoting a particular product, and therefore be able to make better decisions.

Market Basket Analysis (MBA) is a widely studied area within machine learning and aims to identify these positively correlated products based on sale history [*citation necessary?*]. Studies on cannibalization products and how machine learning can be applied to identify them are however much less extensive.

Identifying correlated and cannibalization products requires a large amount of data, machine learning is therefore well suited for solving this problem. By identifying the characteristics of a good model for predicting correlation and the characteristics of a good model for identifying cannibalization products, machine learning can revolutionize the way promotional planning is carried out, increasing the profit of planned promotions for companies all over the world.

As stated above, MBA is a widely studied area and many research papers exist that discuss the best models and algorithms for identifying correlated products [*citation?*]. Our main focus will therefore not be on this part of the research, instead we plan on applying best practice techniques to identify correlation, providing more domain experience.

* See: R. Agrawal Fast algorithms for mining association rules in large databases

More focus will be put on experimenting best ways to identify cannibalization products. Today, cannibalization products are usually identified using statistical models that only look at sale history [*citation?*]. Cannibalization products are however usually products similar to the one being promoted, i.e. if you put Coca Cola on promotion you are likely to see a decrease in the sales of Pepsi during that period [*citation?*]. We would therefore like to explore whether product features can be used to improve current practice.

\* Add discussion about other theory

# Purpose of the study

The purpose statement should provide a specific and accurate summary of the overall pur- pose of the study. Briefly define and delimit the specific area of the research. Incorporate the rationale for the study. A commonly used sentence starts with: “The purpose of this study is . . .”. The purpose should clarify who is anticipated to benefit from the results of your study and how the results may be used.

The purpose of this study is twofold.

* Something about promotional planning (correlation, but mostly cannibalization?)
* Something about the other theory we will come up with

# Review of the literature

The literature review provides the background and context for the research problem. It should establish the need for the research and indicate that the writer is knowledgeable about the area. The literature review:

* + Describes the results of other studies that are closely related to the study being proposed (or reported)
  + Relates a study to the larger, ongoing dialogue in the literature about a topic, filling in gaps and extending prior studies
  + Provides a framework for establishing the importance of the study, as well as a benchmark for comparing the results of a study with other findings
  + “Frames” the problem earlier identified

The literature review should demonstrate to the reader that you have a comprehensive grasp of the field and are aware of important recent substantive and methodological devel- opments. Define the starting point for your study - how will your study refine, revise, or extend what is now known?

In a proposal, the literature review is generally brief and to the point. Select and reference only the more appropriate citations. Make key points clearly and succinctly. Later in your thesis, you will elaborate on this section.

Some reference about cannibalization products being similar products?

Some reference about what model to use for market basket analysis?

???

# Research question and/or Hypotheses

Questions are relevant to descriptive, normative or census type research. (What are rele- vant factors? How many of them are there? Is there a relationship between them?) Hypoth- eses are relevant to theoretical research, and when you state hypotheses the reader is entitled to have an exposition of the theory that lead to them (and the assumptions under- lying the theory).

In general, you should be prepared to interpret any possible outcome with respect to the questions or hypotheses. Try to visualize in your mind tables or other summary devices, which you expect to come out of the research, short of the actual data.

In this project we would like to first examine whether the cannibalized products can be identified by learning their features, using a supervised learning. Secondly we would like to see if a machine learning model using product features, as well as the correlation from the sales history, can give a better idea of the total effect of a promotion than simplistic methods like k-nearest neighbours models that only use information from sale history.

*Hypothesis 1:*

Product features can be used to determine if a product will be cannibalized when another product is promoted.

*Hypothesis 2:*

A machine learning system using sale history and product features can give better results than a simplistic method that only uses information from sale history, for example k-nearest neighbours algorithm, when identifying the k most correlated/cannibalized products.

*Hypothesis 3:*

Something related to the other theory we will add

1. **The Design – Methods and Procedures**

Any research or problem solving requires a systematic approach with methods and procedures. *Indicate the steps you will take to answer every question or to test every hypothesis indicated in the previous section, to solve the problem that you are addressing.* There are several research methods, e.g. design research (Collins 2004, Vaishnavi & Kuechler2004/5), case study (Runeson & Höst 2009, Yin 1994), action research (McKay & Marshall 2001), Survey (Grover 1998), and **experiment** (Basili et al 1986) just to mention a few. Different research methods and procedures require different descriptions.

For example for a survey, it becomes vital to describe sampling and instrumentation. The sampling, i.e. the population and how the sample has been drawn from that, needs to be described to clarify to what extent the findings of a study can be generalized to people or situations. You should also outline the instruments you propose to use (surveys, scales, interview protocols, observation grids). For a case study or a design research, other aspects become vital.

Data collection   
For all studies, you need to have a systematic approach for data collection. Outline the general plan for what data to collect, and how. This may include survey administration procedures, interview or observation procedures.   
Also, provide a general outline of the time schedule you expect to follow.

Data Analysis   
For all studies, you need to have a systematic approach for data analysis. Specify the procedures you will use to analyze your data. If coding procedures are to be used, describe these in reasonable detail. For evaluations, describe the criteria to be used in reasonable detail.

We will start by exploring whether product features can generally be used to determine if a product will be cannibalized when another one is promoted.

For this task, supervised learning methods will be used. The data will be labeled by comparing the forecast and real sale of all products during the promotion of a particular product, the products that have decreased sale will be labeled cannibalization products, whereas other products will be labeled non-cannibalization products. A probabilistic classifier will then be built that learns the product features relevant for predicting the probability distribution of the cannibalization products. Different models, parameters and preprocessing techniques will be explored in order to find the best model [*is this too vague?*].

Next, we will use best practice to create a model that identifies the correlated products. [*will we actually have to do that or can we just go straight to creating the whole model?*]

Using this information, we will then create a comprehensive machine learning system that predicts the total effect of a promotion of a particular product. For this we will use unsupervised learning [*do we have to be more specific? – e.g. what methods we will test and why*]

The objective of the research is to come up with a model identifies the *k* most correlated products and *k* most cannibalized products with a better accuracy than a simple k-nearest neighbors algorithm that detects correlation and cannibalization by looking at sale history.

**Data Collection**

The data for this thesis will be provided by Jysk Canada who have been using the AGR inventory optimization software for planning promotions since 2015. The data consists of Jysk’s product range, various product information and details, sale and stock history and information about planned promotions.  
The data contains information about 2.450.516 distinct products and up to three years of sale and stock history. Of those products, 492.462 products have been put on promotion, with the total amount of promotions being 4.240.008.

Data Analysis: Do we need to discuss something about prior data analysis? Something about the k-nearest neighbors model?

Time schedule:

1. Read the background papers and get familiar with the dataset.
2. Perform a feature selection and determine if product features can be used to identify cannibalization products.
3. Do a market basket analysis using association rules to group highly correlated products, with a real-world dataset in R and SQL2016.
4. Build a content-based recommender engine that identifies correlated and cannibalization products with real-world dataset in R and SQL2016.
5. Evaluate the models against a k-nearest neighbors model.
6. Test the model on a second real-world dataset from retailer.

\* Add other theory

1. **Limitations and Delimitations**

A limitation identifies potential weaknesses of the study. Think about your analysis, the nature of self-report, your instruments, and the sample. Think about threats to external or internal validity that may have been impossible to avoid or minimize and explain these. Delimitation addresses how a study will be narrowed in scope. This is where you explain the things that you are not doing and why you have chosen not to do them. For example, the literature you will not review (and why not), the population you are not studying (and why not), the methodological procedures you will not use (and why not). Limit your delimitations to the things that a reader might reasonably expect you to do (given your topic and problem statement) but that you, for clearly explained reasons, have decided not to do.

????

1. **Significance of the study**

Indicate how your research will refine, revise, or extend existing knowledge in the area under investigation. Note that such refinements, revisions, or extensions may have substantive, theoretical, or methodological significance. Think pragmatically. Most studies have two potential audiences: practitioners and researchers. Think about implications: What implications may the results of the study have on research? What implications may the results of the study have on practice?

It is yet to be studied how machine learning models can be used to analyze the secondary effects of a promotion, so the thesis will dive into a sector that has not been studied to any extent. The research will therefore hopefully be a predecessor for further research that can build upon and refine the knowledge obtained.

This research will on top of that prove incredibly significant for retail practitioners all around the world.

\* Add something about other theory